ENTROPY FORCING: ENTROPY-ROUTED SUBSET SELECTION FOR KV CACHE IN STREAMING VIDEO GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: Qualitative comparison of **Entropy Forcing (Ours)** and Deep Forcing (Yi et al., 2025). For both 30- and 60-second long-video generation, Entropy Forcing preserves long-range temporal consistency and visual quality, while producing more dynamic subjects and backgrounds.

ABSTRACT

Autoregressive (AR) video diffusion has emerged as a promising paradigm for streaming and open-ended long-video synthesis, with generation quality depending heavily on effective KV cache management under a fixed budget. Existing methods typically rely on attention-based importance, selecting KV entries with the highest attention scores for subsequent generation. However, the reliability of this importance signal depends on how distinctly the attention scores separate different KV entries. Specifically, well-separated scores yield clear rankings, whereas clustered scores provide limited discrimination and make selection less reliable. In such cases, existing methods lack an appropriate criterion for selecting a suitable subset of KV entries. To address these limitations, we propose **Entropy Forcing**, a training-free KV cache selection method for AR video diffusion. First, instead of scoring KV entries independently, we introduce a coverage-aware objective that explicitly accounts for redundancy among retained historical representations. Second, we use normalized Rényi-2 entropy as a global routing signal to adaptively determine how much the selection should rely on attention-based importance versus representation coverage. Finally, we exhaustively score and rank candidate KV subsets under the given budget, with parallel evaluation enabling efficient inference. Experiments on long-video generation show that Entropy Forcing improves motion dynamics while maintaining competitive visual quality and temporal consistency compared with recent state-of-the-art methods. Notably, it consistently achieves the best overall average performance across the evaluated metrics.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Recent advances have established autoregressive (AR) video diffusion as a promising framework for streaming and long-horizon video generation (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025; Liu et al., 2025; Yang et al., 2025; Zhu et al., 2026). Unlike bidirectional diffusion models (team, 2024), which denoise over a complete temporal context, AR models generate frames or video chunks sequentially, conditioning each step on previously generated content. This sequential formulation naturally supports streaming generation and enables video synthesis beyond a fixed temporal window. In Transformer-based AR video diffusion, historical information is primarily preserved through the Key-Value (KV) cache, which stores intermediate representations from previous frames for reuse in subsequent attention computation. However, retaining the full history causes both memory consumption and attention cost to grow with sequence length (Xi et al., 2026; Ji et al., 2026; Yesiltepe et al., 2026). Effective KV cache management under a fixed context budget is therefore essential for efficient open-ended long-video generation.

Existing approaches mitigate this challenge by compressing or selectively retaining historical KV representations (Yi et al., 2025; Chen et al., 2026a; Luo et al., 2026; Meng et al., 2026). Among these approaches, importance-aware selection offers a simple and effective way to allocate the limited cache budget to entries likely to benefit future generation. A common strategy derives importance from attention statistics, ranks candidate KV entries by their attention scores, and retains the top-scoring ones (Zhang et al., 2023; Li et al., 2024). This design is intuitive because attention directly reflects the relevance of cached keys to the current queries. When the scores are well separated, the resulting ranking provides a clear basis for distinguishing more relevant entries from less relevant ones.

However, the reliability of attention-based selection depends on how distinctly the attention scores separate different candidates. When the scores are closely clustered, the induced ranking becomes weakly discriminative, causing retention decisions to hinge on small numerical differences. This suggests that the reliance on attention should adapt to the discriminative strength of its score distribution rather than remain fixed. Moreover, attention captures the relevance of historical keys to the current queries but does not account for redundancy and complementarity among the retained value representations. Consequently, a set of individually high-scoring entries may still provide poor coverage of the historical representation space. Accounting for such subset-level interactions calls for joint optimization of the retained cache, rather than treating candidate entries independently.

To address these limitations, we propose **Entropy Forcing**, a training-free KV cache selection method for AR video diffusion. In addition to attention-based importance, we introduce a coverage-aware score over cached value representations to account for redundancy and complementarity among candidate entries. We then use normalized Rényi-2 entropy (Rényi, 1961) to characterize the discriminative strength of the attention scores and adaptively balance attention-based importance with representation coverage. Based on the resulting selection objective, we formulate cache management as a subset selection problem. At each Transformer layer, Entropy Forcing evaluates feasible candidate subsets in parallel under a fixed memory budget and selects the subset with the highest objective value, enabling exact selection without additional training. Figure 2 illustrates the overall pipeline.

We evaluate Entropy Forcing on long-video generation using VBench-Long (Huang et al., 2024). Across both 30-second and 60-second generation settings, Entropy Forcing improves motion dynamics while maintaining competitive visual quality and temporal consistency compared with recent state-of-the-art AR video generation baselines. Overall, the proposed method achieves the highest average performance across the evaluated metrics in both settings.

Our contributions are summarized as follows:

- We identify two limitations of conventional attention-based KV cache selection in AR video diffusion. First, the reliability of the induced ranking depends on the separation among candidate attention scores, providing weak discrimination when many candidates receive similar scores. Second, independent entry-wise scoring ignores redundancy among the retained representations.
- We propose **Entropy Forcing**, a training-free KV cache selection method that extends conventional attention-based selection with a coverage-aware subset objective to explicitly account for redundancy and complementarity among historical representations. We further use Rényi-2 entropy to characterize the discriminative strength of the current attention distribution and adaptively balance attention-based importance and representation coverage.

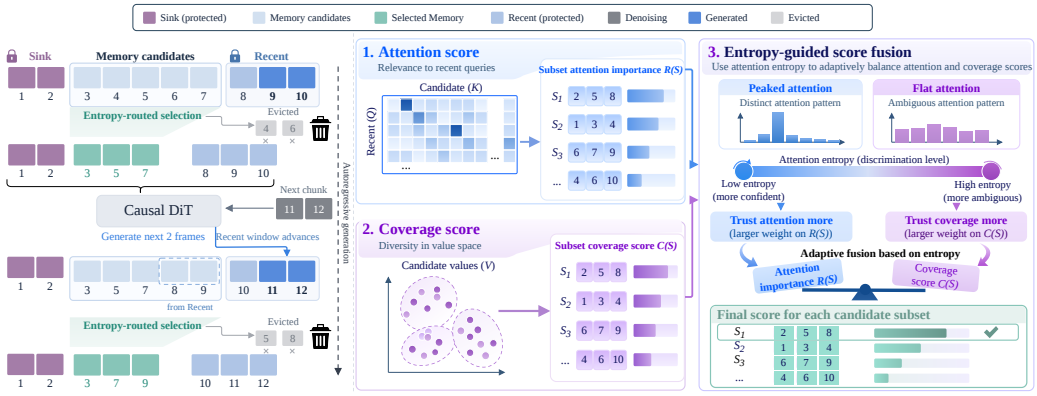


Figure 2: Overview of Entropy Forcing. At each Transformer layer, the KV cache is partitioned into protected Sink and Recent frames and a selectable Memory region. For each candidate Memory subset, we compute (1) an attention-based importance score from current-query attention and (2) a representation coverage score from cached value features. The normalized Rényi-2 entropy of the attention distribution measures its discriminative strength and adaptively balances the two scores: lower entropy places greater weight on attention-based importance, whereas higher entropy increases the contribution of representation coverage. The resulting fused score is used to select the retained Memory subset under the fixed cache budget.

- We formulate cache selection as an entropy-routed subset optimization problem and obtain the exact solution by evaluating all feasible subsets over the bounded Memory candidate set in parallel.
- Experiments demonstrate improved motion dynamics and the best overall average performance across the evaluated metrics while maintaining competitive visual quality and temporal consistency.

2 Related Work

Streaming Autoregressive Video Generation. Recent works combine diffusion modeling with autoregressive prediction to support long-horizon and streaming video generation. Diffusion Forcing (Chen et al., 2024) establishes a general framework that unifies next-token prediction with token-wise diffusion. CausVid (Yin et al., 2025) further converts a pretrained bidirectional diffusion transformer into a causal generator and enables efficient autoregressive inference with KV caching. Self Forcing (Huang et al., 2025) reduces the train–inference mismatch by training the model on its own generated trajectories. Subsequent methods, including Future Forcing, TetherCache, Deep Forcing, and other streaming video generators (Liu et al., 2025; Yang et al., 2025; Zhu et al., 2026), further improve long-term consistency, historical-context utilization, and efficient causal generation. These works primarily study the generation or training paradigm itself, whereas our work focuses on how historical KV entries should be retained under a fixed cache budget during inference.

KV Cache Management for Long-Context Generation. KV cache compression has been widely studied in large language models. StreamingLLM (Xiao et al., 2024) identifies attention sinks and preserves a small set of initial tokens together with recent context, while H_2O (Zhang et al., 2023) retains heavy-hitter tokens according to accumulated attention. Related methods further explore attention-based eviction, adaptive cache budgets, token merging, and representation compression (Liu et al., 2023; Li et al., 2024; Feng et al., 2024). Recent AR video generation methods extend similar ideas to temporal KV management, using attention relevance, recent or future context, temporal diversity, or compressed memory representations to preserve useful historical information (Yi et al., 2025; Chen et al., 2026a; Luo et al., 2026; Ji et al., 2026; Meng et al., 2026; Shi et al., 2026). Most selection-based approaches nevertheless rely heavily on attention scores to rank historical entries. Entropy Forcing instead considers both attention-based relevance and redundancy in the value space, and uses attention entropy to adaptively balance the two signals for subset-level memory selection.

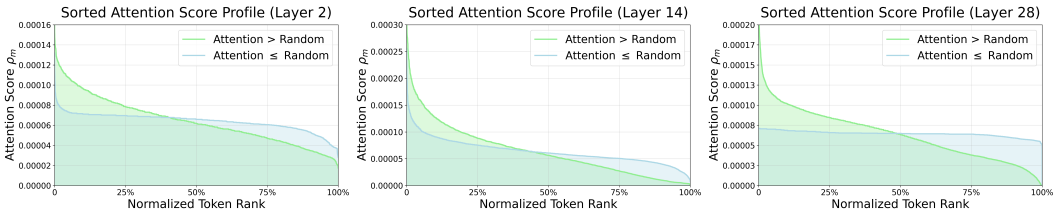


Figure 3: Sorted attention score profiles across different Transformer layers of Wan2.1-T2V-1.3B. Attention Top-B outperforms random selection when the score decay is steep, indicating highly distinguishable token importance. Conversely, this advantage vanishes when the profile becomes flat, as uniform attention scores among candidates render rank-based selection ineffective.

3 Preliminaries

Autoregressive (AR) video diffusion. We consider AR video generation conditioned on a prompt c (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025). Let $\mathbf{x}_{1:T} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ denote a video sequence, where $\mathbf{x}_i \in \mathbb{R}^{N \times d}$ contains the N latent tokens of the i-th frame. The joint distribution is:

$$p(\mathbf{x}_{1:T} \mid c) = \prod_{i=1}^{T} p(\mathbf{x}_i \mid \mathbf{x}_{<i}, c). \tag{1}$$

For a token with hidden representation $\mathbf{h}_n \in \mathbb{R}^d$, its query, key, and value are:

$$\mathbf{q}_n = W_Q \mathbf{h}_n, \qquad \mathbf{k}_n = W_K \mathbf{h}_n, \qquad \mathbf{v}_n = W_V \mathbf{h}_n. \tag{2}$$

The keys and values of previously generated frames are stored in the KV cache and reused during subsequent generation.

Attention-based KV Cache Selection. At each selection state, we divide the historical context into a protected region $\mathcal{P}$, containing Sink and Recent frames, and a selectable Memory region $\mathcal{M}$. Given a Memory budget B, each Transformer layer ℓ retains a subset $\mathcal{S}^{(\ell)} \subseteq \mathcal{M}$ with $|\mathcal{S}^{(\ell)}| = B$.

Before selection, the current queries attend to all historical candidates in $\mathcal{M}$ using standard scaled dot-product attention (Vaswani et al., 2017):

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{Q K_{\mathcal{M}}^{\top}}{\sqrt{d}}\right). \tag{3}$$

For each Memory frame $m \in \mathcal{M}$, let $\mathcal{T}_m$ denote its visual tokens. At Transformer layer ℓ, its frame-level attention score is defined as:

$$p_m^{(\ell)} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{j \in \mathcal{T}_m} \mathbb{E}_{h,\tau}\left[A_{qj}^{(\ell,h,\tau)}\right], \qquad m \in \mathcal{M}, \tag{4}$$

where $\mathcal{Q}$ denotes the query tokens used for cache selection and $\mathbb{E}_{h,\tau}$ denotes averaging over attention heads and denoising steps. Conventional attention-based selection ranks Memory frames according to $p_m^{(\ell)}$ and retains the top-B frames at each layer (Yi et al., 2025; Chen et al., 2026a; Luo et al., 2026).

4 Methodology

4.1 Motivation: Limitations of Attention-Based Selection

Attention-based KV cache selection ranks candidates based on their relevance to the current queries and retains the top-scoring entries under a fixed budget. While this provides a clear selection criterion when candidate scores are well separated, its reliability diminishes as the scores become more closely clustered. Moreover, it does not explicitly account for redundancy among the retained representations. See the next two paragraphs for detailed empirical validation and analysis.

Attention-based selection becomes less reliable as score separation decreases. We compare attention-based selection with random selection under the same cache budget. As shown in Figure 3, attention-based strategy consistently outperforms random selection when candidate attention scores are well separated, as reflected by the steeper decay of the green curve. In contrast, this advantage diminishes when the scores become clustered, as in the middle region of the blue curve, where most candidates receive similar attention scores. Additional experimental settings and more layers results are provided in Appendix A. These observations indicate that the reliability of attention-based importance depends on the discriminative strength of the attention scores. When score separation is small, minor differences in attention provide limited evidence for determining which candidates should be retained. The reliance on attention-based importance should therefore adapt to the degree of separation among the current candidate scores.

Attention-based importance does not account for redundancy within the retained subset. To expose this limitation, we reformulate conventional Attention-based Top-B selection at the subset level. For a candidate subset $\mathcal{S} \subseteq \mathcal{M}$, we define its attention-based importance as the total attention mass retained by the subset:

$$R^{(\ell)}(\mathcal{S}) = \sum_{m \in \mathcal{S}} p_m^{(\ell)}. \tag{5}$$

Crucially, $R^{(\ell)}(\mathcal{S})$ is additive over individual candidates. For any $m \notin \mathcal{S}$,

$$R^{(\ell)}(\mathcal{S} \cup \{m\}) - R^{(\ell)}(\mathcal{S}) = p_m^{(\ell)}, \tag{6}$$

which means that the contribution of candidate m is independent of the candidates already retained in $\mathcal{S}$. Consequently, the objective cannot distinguish between subsets whose members carry complementary information and those containing redundant value representations, allowing multiple high score candidates to encode similar historical information while occupying separate Memory slots.

We further verify that such redundancy occurs in the historical KV cache. Figure 4 shows the pairwise cosine similarity between frame-level value representations across selected Transformer layers. In addition to self-similarity along the anti-diagonal, many off-diagonal frame pairs exhibit high similarity, showing that redundant representations can exist in Memory budget. Additional details are provided in Appendix A.2.

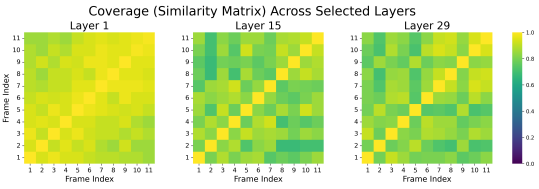


Figure 4: Pairwise similarity of frame-level value representations across selected layers. High similarity away from the anti-diagonal indicates substantial redundancy among historical representations that is not captured by attention-based selection.

Together with the preceding observation, these results reveal two limitations of attention-based cache selection. Its reliability degrades when attention scores provide weak separation among candidates, while its additive objective ignores redundancy within the selected subset. These limitations motivate a selection criterion that accounts for representation coverage while adapting the reliance on attention-based importance to the discriminative strength of the current attention scores.

4.2 Coverage-Aware Subset Selection

To account for redundancy within the retained Memory, we evaluate candidate frames jointly according to how well the selected subset covers the historical value representations. For each candidate $m \in \mathcal{M}$ at Transformer layer ℓ, we obtain a frame-level representation by mean-pooling its cached value tokens,

$$z_m^{(\ell)} = \frac{1}{|\mathcal{T}_m|} \sum_{j \in \mathcal{T}_m} v_j^{(\ell)}, \tag{7}$$

where $\mathcal{T}_m$ denotes the visual tokens belonging to frame m. We then compute the cosine similarity between frame-level representations,

$$\hat{z}_m^{(\ell)} = \frac{z_m^{(\ell)}}{\|z_m^{(\ell)}\|_2}, \qquad s_{ij}^{(\ell)} = \left(\hat{z}_i^{(\ell)}\right)^\top \hat{z}_j^{(\ell)}. \tag{8}$$

A larger $s_{ij}^{(\ell)}$ indicates greater similarity between the historical representations of candidates i and j. Given a retained subset $\mathcal{S} \subseteq \mathcal{M}$, we define its attention-weighted representation coverage score as:

$$C^{(\ell)}(\mathcal{S}) = \sum_{j \in \mathcal{M}} p_j^{(\ell)} \max_{i \in \mathcal{S}} s_{ji}^{(\ell)}. \tag{9}$$

For each historical candidate j, the maximum similarity measures how well its representation is covered by the most similar retained candidate, while $p_j^{(\ell)}$ assigns greater weight to candidates with higher attention-based importance. Unlike the additive attention objective in Eq. (5), the contribution of a candidate under Eq. (9) depends on which representations are already covered by the selected subset. For $m \notin \mathcal{S}$, its marginal coverage gain is:

$$\begin{aligned} \Delta_C^{(\ell)}(m \mid \mathcal{S}) &= C^{(\ell)}(\mathcal{S} \cup \{m\}) - C^{(\ell)}(\mathcal{S}) \\ &= \sum_{j \in \mathcal{M}} p_j^{(\ell)} \max\left(0,\ s_{jm}^{(\ell)} - \max_{i \in \mathcal{S}} s_{ji}^{(\ell)}\right). \end{aligned} \tag{10}$$

Intuitively, a candidate similar to already covered representations yields little additional gain, whereas one covering underrepresented representations contributes more. The objective therefore favors complementary candidates and discourages redundant entries from occupying multiple Memory slots. This formulation follows the facility-location principle for representative subset selection (Lin & Bilmes, 2011); a geometric interpretation is provided in Appendix C.1.

4.3 Entropy-Guided Adaptive Fusion

Representation coverage addresses redundancy within the selected subset, but a fixed balance between attention-based importance and coverage is suboptimal because the reliability of attention scores varies across layers and generation steps. As observed in Section 4.1, attention-based selection becomes less reliable as the attention distribution becomes flatter. We therefore use the entropy of the current attention distribution to characterize its discriminative strength and adaptively balance attention-based importance and representation coverage.

Let $N = |\mathcal{M}|$ denote the number of Memory candidates, and let $\mathbf{p}^{(\ell)}$ denote the attention distribution defined in Eq. (4). We measure its evenness using the normalized Rényi-2 entropy (Rényi, 1961):

$$H^{(\ell)} = -\frac{\log\left(\sum_{m \in \mathcal{M}} \left(p_m^{(\ell)}\right)^2\right)}{\log N}, \qquad H^{(\ell)} \in [0, 1]. \tag{11}$$

Low entropy corresponds to a concentrated distribution in which a small number of candidates receive substantially larger attention scores, whereas high entropy indicates a flatter distribution with weaker score separation. Moreover, the entropy is directly related to the deviation of the attention distribution from uniformity. Let $\mathbf{u} = (1/N, \ldots, 1/N)$ denote the uniform distribution, Eq. (11) yielding:

$$\left\|\mathbf{p}^{(\ell)} - \mathbf{u}\right\|_2^2 = N^{-H^{(\ell)}} - \frac{1}{N}. \tag{12}$$

Thus, for a fixed candidate-set size, increasing $H^{(\ell)}$ corresponds exactly to decreasing the ℓ_2 distance between the attention distribution and the uniform distribution. Rényi-2 entropy therefore provides a continuous measure of how strongly the current attention scores discriminate among candidates. The complete derivation is provided in Appendix B.

We further verify this relationship empirically in Figure 5. For each candidate-set selection strategy, we retain B candidates and compute the resulting attention layer output using the selected subset. We use the output obtained with the full candidate set as the reference and measure the deviation of each selected-subset output from this reference. Taking random selection as the baseline, we define the advantage of Attention Top-B as the difference between the reconstruction errors of random selection and Attention Top-B, denoted by ΔQ. A larger ΔQ therefore indicates a greater advantage of attention-based selection over random selection in preserving the full attention output. We observe a strong negative correlation between entropy and this advantage (Pearson $r = -0.80$, $p \leq 0.001$): as the attention distribution becomes flatter, the advantage of Attention Top-B progressively diminishes.

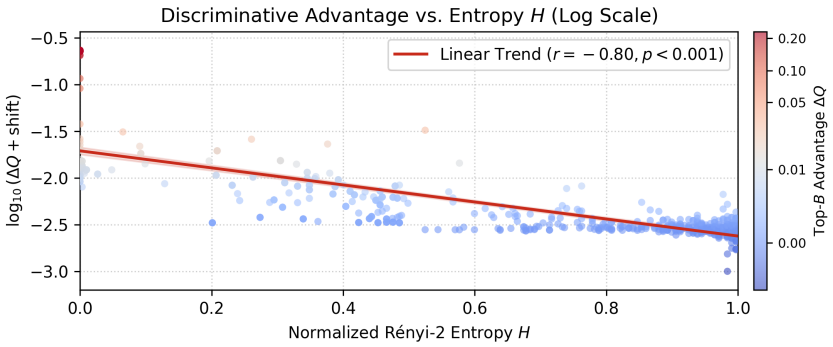


Figure 5: Relationship between the normalized Rényi-2 entropy and the relative advantage of attention-based selection over random selection.

This result is consistent with the observation in Section 4.1 and supports using $H^{(\ell)}$ to modulate the reliance on attention-based importance.

We therefore assign weight $1 - H^{(\ell)}$ to attention-based importance and $H^{(\ell)}$ to representation coverage, yielding the adaptive subset-selection objective:

$$\mathcal{S}^{\star(\ell)} = \arg \max_{\substack{\mathcal{S} \subseteq \mathcal{M} \\ |\mathcal{S}|=B}} \left[\left(1 - H^{(\ell)}\right) R^{(\ell)}(\mathcal{S}) + H^{(\ell)} C^{(\ell)}(\mathcal{S}) \right]. \tag{13}$$

When the attention distribution is concentrated, $H^{(\ell)}$ is small, and selection relies primarily on attention-based importance. As the distribution becomes flatter, $H^{(\ell)}$ increases, reducing the reliance on weakly discriminative attention scores and placing greater emphasis on representation coverage to avoid retaining redundant candidates. The resulting objective provides a continuous, parameter-free interpolation between the two criteria.

Although the resulting objective is combinatorial, it can be solved efficiently because the Memory candidate set is bounded. We enumerate all feasible size-B subsets and evaluate them jointly in parallel, rather than scoring each subset sequentially. All quantities required for subset evaluation, including the normalized attention scores and pairwise representation similarities, are computed once before enumeration and shared across all candidate subsets. Consequently, evaluating a subset requires only lightweight indexing and reduction operations, allowing us to obtain the exact optimum with low additional overhead. The complete selection procedure is provided in Algorithm 1 in Appendix C.2.

5 Experiment

5.1 Main Experiment

Implementation Details. We implement our Entropy Forcing on the Wan2.1-T2V-1.3B model (Wang et al., 2025) and evaluate it against several recent state-of-the-art autoregressive video diffusion baselines, including CausVid (Yin et al., 2025), Self Forcing (Huang et al., 2025), Rolling Forcing (Liu et al., 2025), LongLive (Yang et al., 2025), Deep Forcing (Yi et al., 2025), MemRoPE (Kim et al., 2026) and Pyramid Forcing (Chen et al., 2026b). We use $S = 4$ Sink frames and $R = 4$ Recent frames as protected regions, together with a retained Memory frames of $B = 8$, resulting in a total KV cache budget of 16 frames. For each chunk, we decode 3 frames. Entropy Forcing is applied only to the selectable Memory region, while the Sink and Recent frames are always preserved.

Evaluation Settings. We evaluate long-video generation on VBench-Long (Huang et al., 2024) using 128 prompts from MovieGen (team, 2024), following the same prompt construction protocol as Deep Forcing (Yi et al., 2025). We evaluate both 30- and 60-second video generation.

Quantitative Results. Table 1 reports the quantitative results on VBench-Long. Entropy Forcing consistently improves Dynamic Degree over all compared methods, reaching 56.30 and 53.65 for 30-second and 60-second generation, respectively. Compared with our reproduced Deep Forcing baseline,

Table 1: Quantitative comparison on long-video generation. We evaluate Entropy Forcing against open-source autoregressive video diffusion baselines on 30- and 60-second videos using VBench-Long. $*$ denotes our reproduction of Deep Forcing (Yi et al., 2025), MemRoPE (Kim et al., 2026) and Pyramid Forcing (Chen et al., 2026b) while † denotes results reported in the Deep Forcing paper.

Method	Dynamic Degree ↑	Motion Smoothness ↑	Overall Consistency ↑	Imaging Quality ↑	Aesthetic Quality ↑	Subject Consistency ↑	Background Consistency ↑	Avg. ↑
30 seconds								
Rolling Forcing†	30.71	98.75	20.99	70.58	60.24	**98.12**	96.91	68.04
LongLive†	45.55	98.76	20.16	69.07	61.51	97.97	96.83	69.98
CausVid†	47.21	98.08	19.15	66.36	59.77	97.92	96.77	69.32
Results on Self Forcing (Huang et al., 2025)								
Self Forcing†	36.62	98.63	20.50	68.58	59.44	97.34	96.47	68.23
+Deep Forcing*	52.90	98.66	**25.31**	67.66	61.10	97.28	96.56	71.35
+MemRoPE*	55.64	98.03	22.01	68.70	58.45	96.30	95.64	70.68
+Pyramid Forcing*	53.03	98.16	24.94	69.69	61.05	96.92	95.94	71.39
+Entropy Forcing	**56.30**	**98.78**	25.10	**70.91**	**62.18**	97.79	**97.22**	**72.61**
60 seconds								
Rolling Forcing†	31.35	98.69	20.64	70.25	59.75	**97.97**	96.76	67.92
LongLive†	43.49	**98.75**	20.29	69.11	61.29	97.85	96.74	69.65
CausVid†	46.44	98.09	18.78	65.84	59.42	97.81	96.75	69.02
Results on Self Forcing (Huang et al., 2025)								
Self Forcing†	31.98	98.21	18.63	66.33	56.45	96.82	96.31	66.39
+Deep Forcing*	52.65	97.76	**25.18**	65.99	60.42	97.22	96.17	70.77
+MemRoPE*	53.08	98.07	20.96	68.72	58.48	96.36	95.60	70.18
+Pyramid Forcing*	52.23	98.13	24.01	69.71	60.95	96.86	95.90	71.11
+Entropy Forcing	**53.65**	98.35	24.55	**70.42**	**61.51**	97.50	**96.90**	**71.84**

Entropy Forcing also significantly improves Imaging Quality, Aesthetic Quality and Background Consistency. For the remaining metrics, Entropy Forcing also maintains competitive performance. Its Overall Consistency remains close to Deep Forcing while exceeding the other reported baselines, and its Motion Smoothness and Subject Consistency remain comparable to existing methods. Overall, it achieves the highest average score in both two settings.

Qualitative Results. Figure 1 presents qualitative comparisons with Deep Forcing under the same cache budget, while Figures 11 and 12 provide additional comparison examples. Entropy Forcing produces more dynamic subject and background motion while maintaining stable visual quality over extended generation. These observations are consistent with the quantitative results, where the main improvement appears in Dynamic Degree without a substantial degradation in the remaining metrics.

5.2 Ablation Study

5.2.1 verifies the effectiveness of both attention-based relevance and value-space coverage, and demonstrates that entropy-guided routing effectively balances the two criteria.

5.2.2 shows that our method incurs negligible additional subset-selection overhead.

5.2.1 Ablation on Entropy-Routed Selection

We examine whether entropy-based routing benefits from combining attention-based importance and value-space representation coverage rather than relying on either criterion alone. We compare three variants under the same Memory budget. *Attention-only* maximizes $R^{(\ell)}(\mathcal{S})$, corresponding to standard Attention Top-B selection, while *Coverage-only* maximizes $C^{(\ell)}(\mathcal{S})$. *Entropy-Routed* uses the complete objective in Equation 13, adaptively balancing the two criteria. Results are shown in Table 2. We evaluate the three variants using CLIP-text similarity (Hessel et al., 2021) and temporal consistency. "CLIP-T" denotes the CLIP-text similarity between the generated content and its text prompt, while "Temp-Consist." denotes the mean cosine similarity between CLIP image features of consecutive frames. All scores are averaged over 20 prompts under the same 30-second setting as the main experiments and evaluated at the 45th chunk.

Entropy-Routed achieves a CLIP-T score of 0.3635 and a temporal consistency score of 0.9836, outperforming both Attention-only and Coverage-only. These results indicate that neither attention-based importance nor value-space coverage alone is sufficient to provide the best selection. Instead,

Table 2: Comparison on different memory selection method.

Method	CLIP-T ↑	Temp-Consist. ↑
Attention-only	0.3600	0.9817
Coverage-only	0.3624	0.9790
Entropy-Routed	**0.3635**	**0.9836**

Table 3: Runtime overhead of entropy-routed Memory selection on a single NVIDIA RTX 4090 GPU.

Video Length	Selection Time (s)	Total Time (s)	Overhead (%)
30s	0.0178	115.09	0.0155
60s	0.0352	227.27	0.0155
120s	0.0697	465.52	0.0150

adapting their relative contributions according to the evenness of the Memory attention distribution improves both semantic alignment and temporal consistency.

We do not report frame-level reconstruction metrics such as LPIPS or SSIM against the full-cache. Because different KV cache selection can lead to different but valid generation trajectories, pixel-level or frame-aligned reconstruction error is not an appropriate measure of cache selection quality.

For qualitative comparison, we additionally include a full-cache reference, which retains the complete historical KV cache without Memory truncation. This reference is used only to visualize the generation trajectory without the bounded-cache constraint and is not treated as ground-truth.

Figure 6 shows the qualitative comparison. The Entropy-Routed variant (the second row, 2 metrics) produces a trajectory visually closer to the full-cache reference than either single-criterion variant. In particular, the morphology of the monkey's arm is better preserved, suggesting that adaptively balancing attention-based importance and representation coverage yields better selection than relying on either criterion alone.

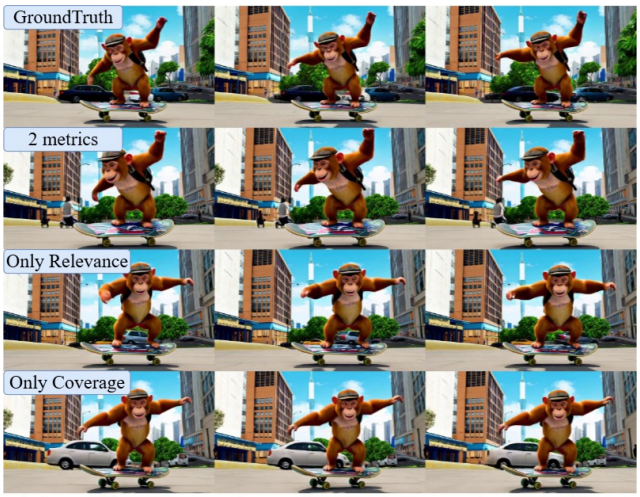


Figure 6: Qualitative comparison of selection strategies. Full-cache reference retains the complete historical KV cache, while all other variants use the same bounded budget. Entropy-Routed better preserves the visual trajectory, including the monkey's arm morphology.

5.2.2 Computation Overhead

As described in Section 4.3, Eq. (13) evaluates all feasible combinations to obtain the optimal subset under the entropy-routed criterion. This exhaustive search remains computationally inexpensive because causal video generation typically operates with a very small context window. In our setting and Deep Forcing, the cache budget is 16 frames with a chunk size of 3 frames. Deep Forcing uses 10 sink and 4 recent frames, leaving 5 candidates for selecting 2 frames, while our method uses 4 sink and 4 recent frames, leaving 11 candidates for selecting 8 frames. Thus, the numbers of evaluated combinations are only $\binom{5}{2} = 10$ and $\binom{11}{8} = 165$, respectively.

Table 3 reports the selection overhead measured on a single NVIDIA RTX 4090 GPU. The selection time is negligible, accounting for less than 0.02% of the total generation time.

6 Conclusion

We presented **Entropy Forcing**, a training-free KV cache selection method for autoregressive video diffusion under a bounded context budget. The method combines attention-based relevance with subset-level representation coverage and uses the evenness of the attention distribution to adaptively balance the two criteria during inference. Experiments on long-video generation demonstrate consistent improvements in both visual quality and temporal consistency over recent state-of-the-art KV cache selection methods, while introducing negligible additional inference overhead. These results suggest that effective KV cache management should consider not only the importance of individual historical entries, but also the representativeness and redundancy of the retained context.

AI USE STATEMENT

Generative AI tools were used solely for language editing, including improving grammar, wording, and readability of the manuscript. They were not used for research ideation, methodology design, experiment implementation, or interpretation of results. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on improving KV cache selection for autoregressive video generation and does not involve human subjects or the collection of private or sensitive data. Our method does not introduce new generative capabilities beyond those of the underlying pre-trained video generation models. Potential risks associated with generated video content are therefore inherited from the base models.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model configuration, cache settings, generation protocol, evaluation metrics, and ablation settings in the main paper and appendix. The complete Entropy Forcing selection procedure is summarized in Algorithm 1, together with the implementation details required to reproduce the experiments. Code will be released upon publication.

REFERENCES

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10–15, 2024*, pp. 24081–24125. Curran Associates, Inc., 2024.

Hanmo Chen, Chenghao Xu, Xu Yang, Xuan Chen, and Cheng Deng. Past- and future-informed KV cache policy with salience estimation in autoregressive video diffusion. *CoRR*, abs/2601.21896, 2026a.

Jiayu Chen, Junbei Tang, Wenbiao Zhao, Maoliang Li, Jiayi Luo, Zihao Zheng, Jiawei Yang, Guojie Luo, and Xiang Chen. Pyramid forcing: Head-aware pyramid KV cache policy for high-quality long video generation. *CoRR*, abs/2605.13111, 2026b.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-kv: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. *CoRR*, abs/2407.11550, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 7514–7528. Association for Computational Linguistics, 2021.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, pp. 167283–167308. Curran Associates, Inc., 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 21807–21818. IEEE, 2024.

Yicheng Ji, Zhizhou Zhong, Jun Zhang, Qin Yang, XiTai Jin, Ying Qin, Wenhan Luo, Shuiyang Mao, Wei Liu, and Huan Li. Forcing-kv: Hybrid KV cache compression for efficient autoregressive video diffusion models. *CoRR*, abs/2605.09681, 2026.

Youngrae Kim, Qixin Hu, C.-C. Jay Kuo, and Peter A. Beerel. Memrope: Training-free infinite video generation via evolving memory tokens. In *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part LI*, volume 17051 of *Lecture Notes in Computer Science*, pp. 244–261. Springer, 2026.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 510–520. Association for Computational Linguistics, 2011.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *CoRR*, abs/2509.25161, 2025. doi: 10.48550/ARXIV.2509.25161.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. *CoRR*, abs/2305.17118, 2023.

Jiayi Luo, Qiyan Liu, Tengyang Wang, Junhao Liu, Jiayu Chen, Cong Wang, Hanxin Zhu, Chen Gao, Xiaobin Hu, Qingyun Sun, and Zhibo Chen. Future forcing: Future-aware training-free KV cache policy for autoregressive video generation. *CoRR*, abs/2605.30083, 2026.

Yu Meng, Xiangyang Luo, Letian Li, Wenyuan Jiang, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. Tethercache: Stabilizing autoregressive long-form video generation with gated recall and trusted alignment. *CoRR*, abs/2606.13035, 2026.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Shuwei Shi, Zhen Li, Muyao Niu, Chuanhao Li, Bo Zheng, Kaipeng Zhang, and Yinqiang Zheng. Surprise forcing: What to remember, when to skip in long video generation. *CoRR*, abs/2607.18436, 2026.

The Movie Gen team. Movie gen: A cast of media foundation models. *CoRR*, abs/2410.13720, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, 2025.

Haocheng Xi, Shuo Yang, Yilong Zhao, Muyang Li, Han Cai, Xingyang Li, Yujun Lin, Zhuoyang Zhang, Jintao Zhang, Xiuyu Li, Zhiying Xu, Jun Wu, Chenfeng Xu, Ion Stoica, Song Han, and Kurt Keutzer. Quant videogen: Auto-regressive long video generation via 2-bit KV-cache quantization. *CoRR*, abs/2602.02958, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net, 2024.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *CoRR*, abs/2509.22622, 2025.

Hidir Yesiltepe, Jiazhen Hu, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, Hoda Eldardiry, and Pinar Yanardag. VideoMLA: Low-rank latent KV cache for minute-scale autoregressive video diffusion. *CoRR*, abs/2605.30351, 2026.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *CoRR*, abs/2512.05081, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11–15, 2025*, pp. 22963–22974. Computer Vision Foundation / IEEE, 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *CoRR*, abs/2602.02214, 2026. doi: 10.48550/ARXIV.2602.02214.

A Motivation Experiment Details

A.1 Attention score analysis

For the comparison in Figure 3, we carefully control the generation process so that the two selection policies differ only in their choice of Memory frames. Specifically, Attention Top-B selection and random selection start from the same generated prefix, protected cache, Memory candidate set, and generation state. Both policies use the same Memory budget B, diffusion noise, sampling schedule, guidance settings, and all other generation parameters. After making a single Memory selection decision, we generate only the immediately following chunk, thereby isolating the effect of this individual selection decision from subsequent cache updates.

For random selection, we repeat the selection with 10 different random seeds and report the averaged result to reduce the variance introduced by random sampling. To construct the representative cases, we evaluate the two selection policies on 10 prompts and identify the case with the largest advantage of Attention Top-B over random selection and the case with the largest advantage of random selection over Attention Top-B. Here, the selection advantage is measured by the SSIM between the generated chunk and its corresponding ground-truth chunk frames. The resulting two cases represent, respectively, scenarios where attention scores are well separated and where they are closely clustered.

Visualization results across additional Transformer layers are shown in Figure 7.

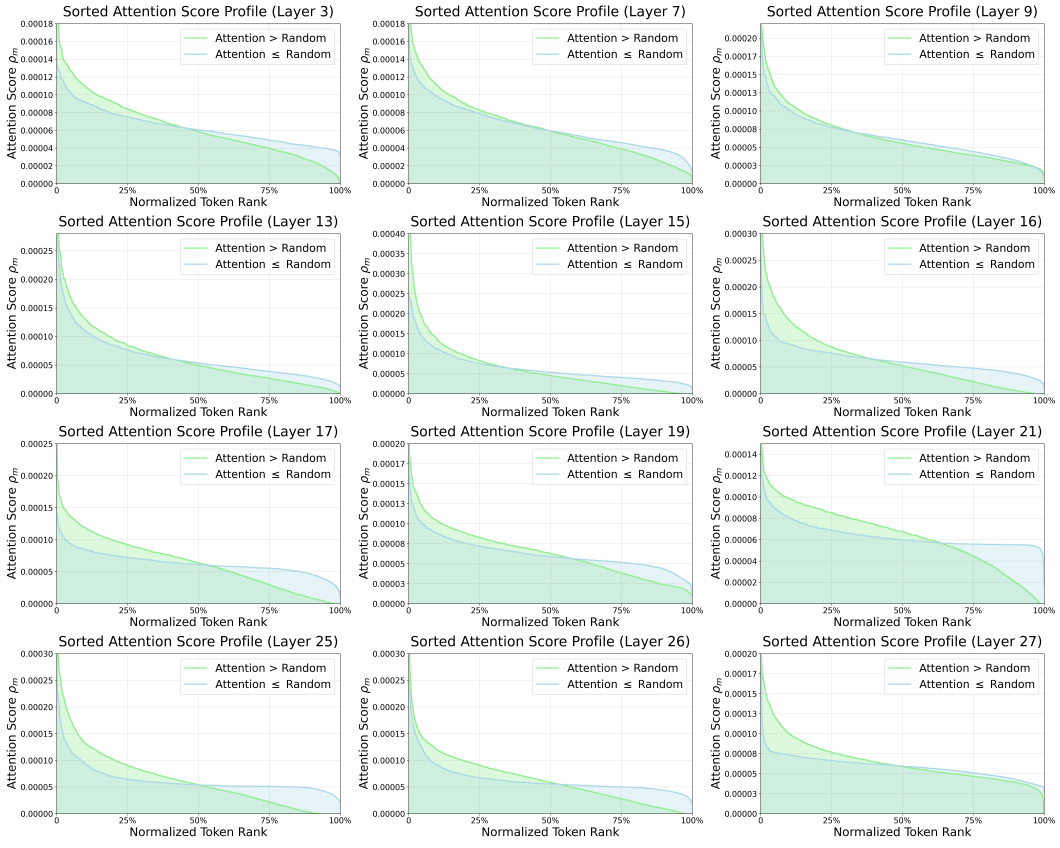


Figure 7: Sorted attention score profiles across additional Transformer layers of Wan2.1-T2V-1.3B, complementing the results presented in the main paper.

A.2 Historical representation redundancy

Figure 4 examines whether redundant historical representations are present in the Memory cache. For each Memory frame, we construct the frame-level value representation defined in Eq. (7) and

normalize it as in Eq. (8). We then compute the pairwise cosine similarity between Memory frames. High-similarity frame pairs indicate that multiple cached frames encode similar historical representations. The observed similarity structure verifies that substantial redundancy can exist among historical Memory frames, motivating the coverage-aware objective introduced in Section 4.2.

B Derivation of the Rényi-2 Entropy Criterion

Section 4.3 uses normalized Rényi-2 entropy to characterize how close the Memory attention distribution is to uniform. Here, we provide the full derivation of the relation used in Eq. (12).

At Transformer layer ℓ, let

$$\mathbf{p}^{(\ell)} = \left(p_1^{(\ell)}, \ldots, p_N^{(\ell)}\right), \qquad \sum_{m=1}^{N} p_m^{(\ell)} = 1, \tag{14}$$

denote the normalized attention distribution over the N Memory candidates.

The normalized Rényi-2 entropy defined in Eq. (11) is

$$H^{(\ell)} = -\frac{\log\left(\sum_{m=1}^{N}\left(p_m^{(\ell)}\right)^2\right)}{\log N}. \tag{15}$$

Rearranging gives

$$\sum_{m=1}^{N}\left(p_m^{(\ell)}\right)^2 = N^{-H^{(\ell)}}. \tag{16}$$

Let

$$\mathbf{u} = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right) \tag{17}$$

denote the uniform distribution over the same candidate set. The squared Euclidean distance between $\mathbf{p}^{(\ell)}$ and $\mathbf{u}$ is

$$\left\|\mathbf{p}^{(\ell)} - \mathbf{u}\right\|_2^2 = \sum_{m=1}^{N}\left(p_m^{(\ell)} - \frac{1}{N}\right)^2 \tag{18}$$

$$= \sum_{m=1}^{N}\left(p_m^{(\ell)}\right)^2 - \frac{2}{N}\sum_{m=1}^{N} p_m^{(\ell)} + \frac{N}{N^2} \tag{19}$$

$$= \sum_{m=1}^{N}\left(p_m^{(\ell)}\right)^2 - \frac{1}{N}. \tag{20}$$

Substituting Eq. (16) yields

$$\left\|\mathbf{p}^{(\ell)} - \mathbf{u}\right\|_2^2 = N^{-H^{(\ell)}} - \frac{1}{N}. \tag{21}$$

Equation Eq. (21) is an exact identity. For a fixed number of Memory candidates, a larger $H^{(\ell)}$ therefore corresponds exactly to a smaller deviation from the uniform distribution. In particular,

$$H^{(\ell)} = 1 \quad \Longleftrightarrow \quad \mathbf{p}^{(\ell)} = \mathbf{u}. \tag{22}$$

At the other extreme, $H^{(\ell)} = 0$ is obtained when all attention mass is concentrated on a single candidate. Thus, $H^{(\ell)}$ provides a normalized measure of the concentration of the attention distribution, with larger values indicating weaker score separation.

The normalization by $\log N$ maps the entropy to the common range $[0, 1]$ regardless of the current Memory candidate-set size. This allows the same routing interpretation to be used when the number of selectable Memory candidates changes.

C Supplementary Method Details

C.1 Geometric Interpretation of Representation Coverage

The representation-coverage objective defined in Eq. (9) is

$$C^{(\ell)}(\mathcal{S}) = \sum_{j \in \mathcal{M}} p_j^{(\ell)} \max_{i \in \mathcal{S}} s_{ji}^{(\ell)}. \tag{23}$$

Because the normalized frame representations $\widehat{z}_i^{(\ell)}$ and $\widehat{z}_j^{(\ell)}$ are unit vectors, their squared Euclidean distance satisfies

$$\left\| \widehat{z}_j^{(\ell)} - \widehat{z}_i^{(\ell)} \right\|_2^2 = \left\| \widehat{z}_j^{(\ell)} \right\|_2^2 + \left\| \widehat{z}_i^{(\ell)} \right\|_2^2 - 2 \left(\widehat{z}_j^{(\ell)} \right)^\top \widehat{z}_i^{(\ell)} \tag{24}$$

$$= 2 - 2 s_{ji}^{(\ell)} \tag{25}$$

$$= 2 \left(1 - s_{ji}^{(\ell)} \right). \tag{26}$$

Since

$$\sum_{j \in \mathcal{M}} p_j^{(\ell)} = 1, \tag{27}$$

we have

$$1 - C^{(\ell)}(\mathcal{S}) = \sum_{j \in \mathcal{M}} p_j^{(\ell)} \left(1 - \max_{i \in \mathcal{S}} s_{ji}^{(\ell)} \right) \tag{28}$$

$$= \sum_{j \in \mathcal{M}} p_j^{(\ell)} \min_{i \in \mathcal{S}} \left(1 - s_{ji}^{(\ell)} \right) \tag{29}$$

$$= \frac{1}{2} \sum_{j \in \mathcal{M}} p_j^{(\ell)} \min_{i \in \mathcal{S}} \left\| \widehat{z}_j^{(\ell)} - \widehat{z}_i^{(\ell)} \right\|_2^2. \tag{30}$$

Therefore, maximizing $C^{(\ell)}(\mathcal{S})$ is equivalent to minimizing the attention-weighted representation distortion between each Memory candidate and its closest retained representative. If a region of the value space is already represented by the selected subset, retaining another highly similar frame provides little additional coverage. The objective therefore naturally discourages redundant frames from occupying multiple positions in the limited Memory budget.

This property distinguishes representation coverage from the additive attention objective in Eq. (5). Under attention-based importance, the contribution of candidate m remains $p_m^{(\ell)}$ regardless of the current subset. Under representation coverage, in contrast, the contribution of a candidate depends on which historical representations have already been covered. Highly redundant candidates can therefore have small marginal utility even when their individual attention scores are large.

C.2 Complete Selection Procedure

Algorithm 1 summarizes the complete inference-time Memory selection procedure of Entropy Forcing. The procedure is performed independently at each Transformer layer ℓ, allowing different layers to retain different Memory subsets. The selected subset $\mathcal{S}^{\star(\ell)}$ replaces the selectable Memory region at layer ℓ, while the protected Sink and Recent regions remain unchanged.

Algorithm 1 Entropy-Routed Exact Subset Selection at Layer ℓ

Require: Memory candidate set $\mathcal{M}$, frame-level attention scores $\{\rho_m^{(\ell)}\}_{m\in\mathcal{M}}$, cached value tokens $\{v_j^{(\ell)}\}$, frame-token sets $\{\mathcal{T}_m\}_{m\in\mathcal{M}}$, Memory budget B

Ensure: Selected Memory subset $\mathcal{S}^{\star(\ell)}$

Normalize the frame-level attention scores: $p_m^{(\ell)} = \dfrac{\rho_m^{(\ell)}}{\sum_{j\in\mathcal{M}} \rho_j^{(\ell)}}, \quad m \in \mathcal{M}.$

Compute normalized Rényi-2 entropy: $H^{(\ell)} = -\dfrac{\log\left(\sum_{m\in\mathcal{M}} (p_m^{(\ell)})^2\right)}{\log|\mathcal{M}|}.$

Mean-pool the value tokens of each Memory frame: $z_m^{(\ell)} = \dfrac{1}{|\mathcal{T}_m|} \sum_{j\in\mathcal{T}_m} v_j^{(\ell)}.$

Compute pairwise cosine similarities: $s_{ij}^{(\ell)} = \dfrac{(z_i^{(\ell)})^\top z_j^{(\ell)}}{\|z_i^{(\ell)}\|_2 \|z_j^{(\ell)}\|_2}.$

$\mathcal{S}^{\star(\ell)} \leftarrow \varnothing$; $\mathrm{bestScore} \leftarrow -\infty$

for each $\mathcal{S} \subseteq \mathcal{M}$ with $|\mathcal{S}| = B$ **do**

  Compute attention-based importance: $R^{(\ell)}(\mathcal{S}) = \sum_{m\in\mathcal{S}} p_m^{(\ell)}.$

  Compute representation coverage: $C^{(\ell)}(\mathcal{S}) = \sum_{j\in\mathcal{M}} p_j^{(\ell)} \max_{i\in\mathcal{S}} s_{ji}^{(\ell)}.$

  Compute the entropy-routed score: $\mathrm{score} \leftarrow (1 - H^{(\ell)})R^{(\ell)}(\mathcal{S}) + H^{(\ell)}C^{(\ell)}(\mathcal{S}).$

  if $\mathrm{score} > \mathrm{bestScore}$ **then**

    $\mathrm{bestScore} \leftarrow \mathrm{score}$

    $\mathcal{S}^{\star(\ell)} \leftarrow \mathcal{S}$

  end if

end for

return $\mathcal{S}^{\star(\ell)}$

D VISUALIZATION EXAMPLES

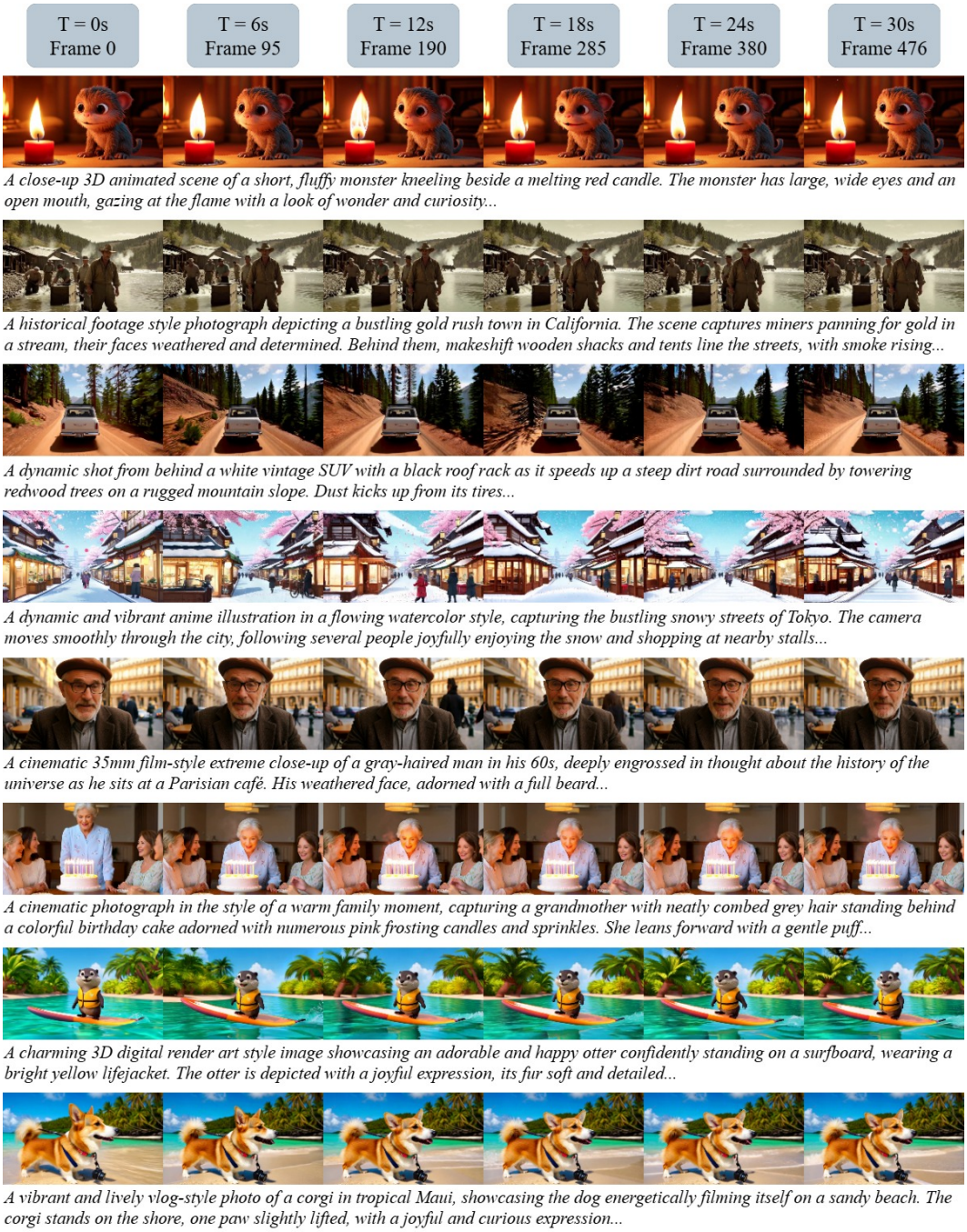


Figure 8: Qualitative visualization of 30-second videos generated by **Entropy Forcing (Ours)** on WAN2.1-T2V-1.3B model using prompts from the MovieGen (team, 2024).

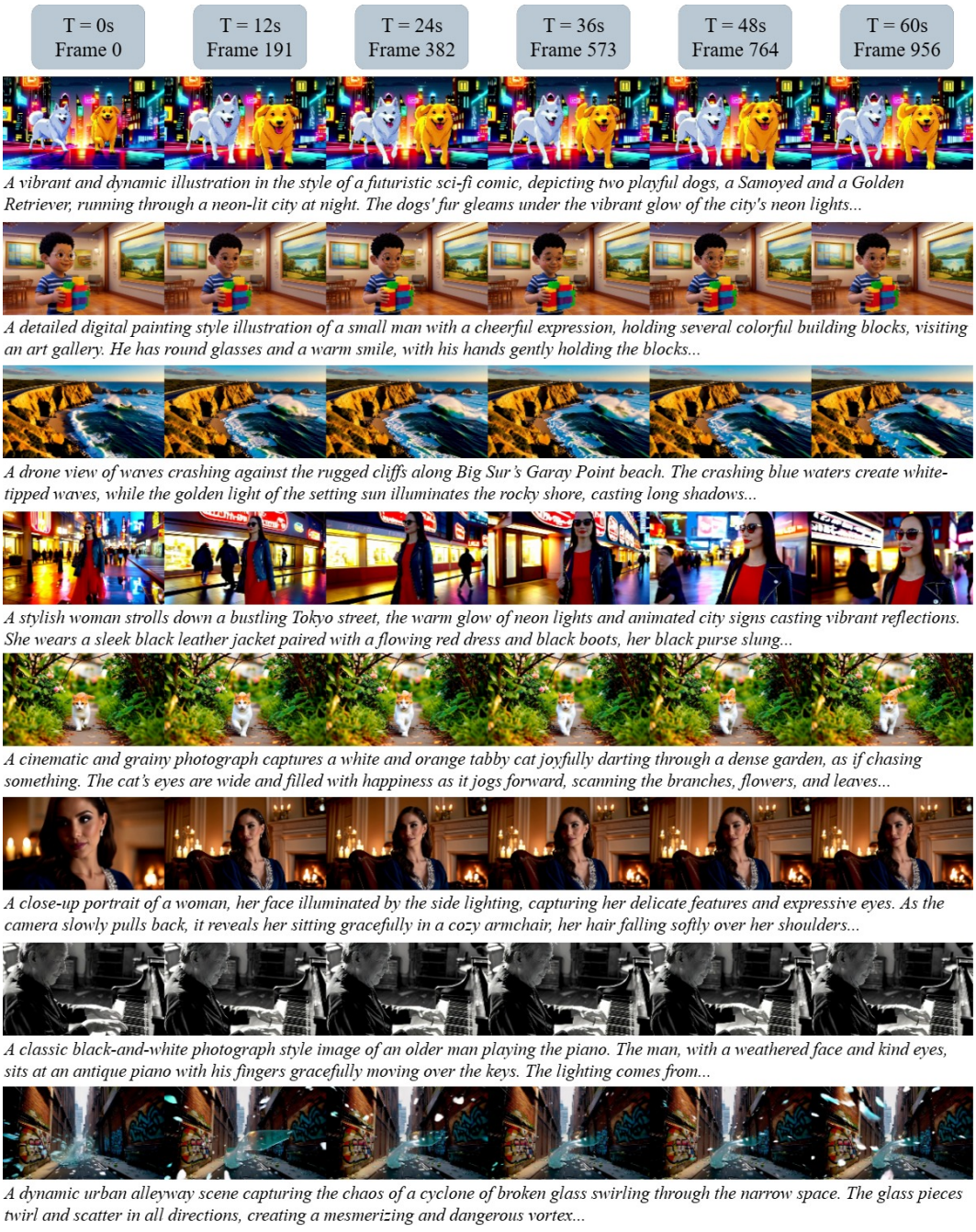


Figure 9: Qualitative visualization of 60-second videos generated by **Entropy Forcing (Ours)** on WAN2.1-T2V-1.3B model using prompts from the MovieGen (team, 2024).



Figure 10: Qualitative visualization of 120-second videos generated by **Entropy Forcing (Ours)** on WAN2.1-T2V-1.3B model using prompts from the MovieGen (team, 2024).

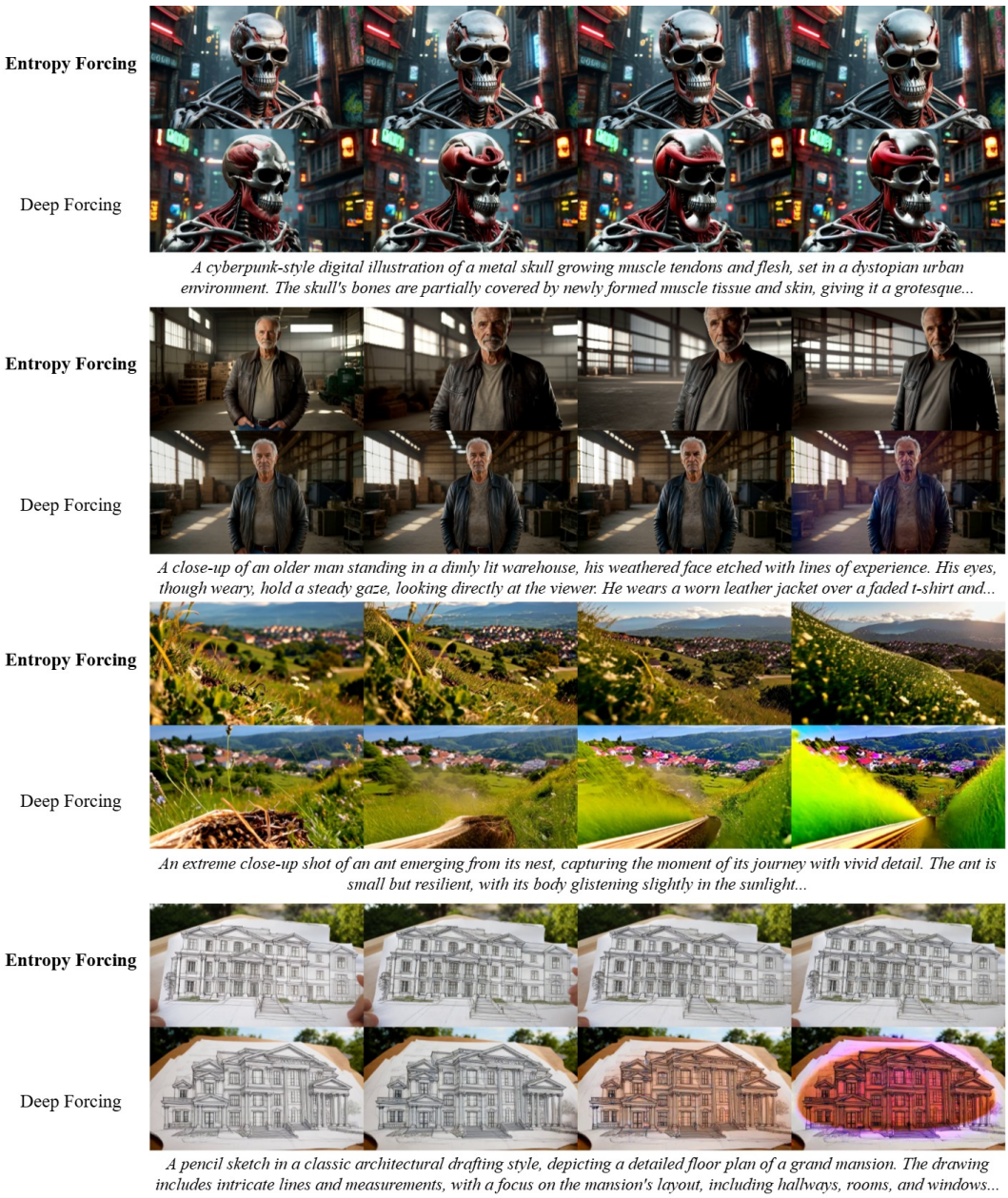


Figure 11: Qualitative comparison between **Entropy Forcing (Ours)** and Deep Forcing (Yi et al., 2025) for 30-second video generation on the MovieGen prompts.



Figure 12: Qualitative comparison between **Entropy Forcing (Ours)** and Deep Forcing (Yi et al., 2025) for 60-second video generation on the MovieGen prompts.